X-ECG: Self-Supervised Explainable Foundation Model for Electrocardiogram

Anonymous ACL submission

Abstract

Electrocardiography (ECG) is widely used for cardiac health evaluation, yet many machine learning methods for ECG analysis lack explainability. We introduce **X-ECG**, a self-supervised explainable ECG foundation model that learns to focus on clinically relevant regions without manual attention annotations. Our key component is Clinically-Guided Attention Localization, which generates attention pseudo-labels on-the-fly using rule-based clinical knowledge and supervises model attention toward these regions via KL divergence loss. This enables the model to highlight abnormal regions contributing to predictions without manual annotations, while improving arrhythmia classification and report generation performance. Experiments show that X-ECG achieves state-of-the-art anomaly localization while achieving state-of-the-art arrhythmia classification and report generation performance.

1 Introduction

Current machine learning models (Zhao et al., 2025; Nguyen et al., 2025; Na et al., 2024; McKeen et al., 2025) for ECG analysis operate as black-boxes: they classify cardiac conditions but cannot show clinicians *where* in the waveform the abnormality occurs. This lack of explainability makes these models difficult to trust in clinical settings, where physicians need to verify model predictions before acting on them.

The fundamental challenge is the absence of attention annotations. Existing ECG datasets such as MIMIC-IV-ECG and PTB-XL contain waveforms with disease labels, but lack annotations indicating which regions of the signal are abnormal. Without such supervision, models cannot learn to localize anomalies. While anomaly detection methods (Bui et al., 2024; Jiang et al., 2023a, 2024) can highlight irregular beats, they do not provide specific diagnoses associated with those anomalies.

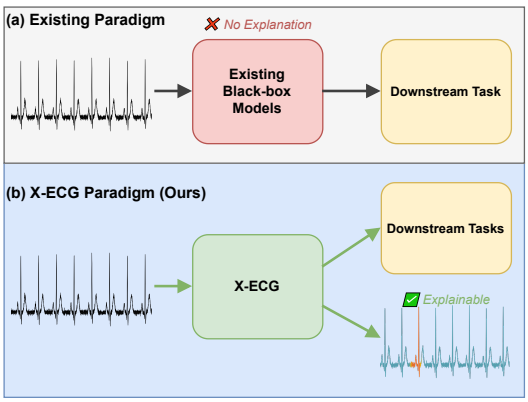


Figure 1: Comparison of (a) existing ECG classification models that act as black-boxes, outputting predictions without explanations, versus (b) X-ECG, which highlights the specific ECG waves contributing to each diagnosis. This enables clinicians to verify model reasoning in seconds rather than re-reading the entire waveform.

We propose a simple insight: established clinical knowledge already defines what constitutes an abnormal ECG pattern. For example, a PR interval exceeding 200ms indicates first-degree AV block, and ST elevation greater than 0.1mV suggests myocardial injury. By encoding these clinical heuristics into rule-based algorithms, we can automatically generate attention pseudo-labels without manual annotation. This transforms the explainability problem from a labeling bottleneck into an engineering task.

Building on this insight, we introduce X-ECG, a self-supervised explainable ECG foundation model. As illustrated in Figure 1, our key component is Clinically-Guided Attention Localization (CGAL), which automatically generates attention pseudo-labels on-the-fly using rule-based clinical heuristics and supervises the model's attention toward these clinically meaningful regions via a Kullback–Leibler (KL) divergence loss. For instance, when X-ECG predicts atrial fibrillation, it highlights the irregular P-wave patterns in Lead II, allowing

the cardiologist to immediately verify the diagnosis. Experiments demonstrate that X-ECG achieves state-of-the-art anomaly localization with 84% AUROC, outperforming existing methods including those specifically designed for this task, while achieving state-of-the-art downstream performance.

Our main contributions are:

- We propose CGAL, a self-supervised mechanism that generates attention pseudo-labels from clinical heuristics and guides model attention toward diagnostically relevant regions, enabling explainable ECG models without manual annotations.
- We demonstrate that this self-supervised attention-guiding approach achieves 84% AUROC on anomaly localization while maintaining strong performance in arrhythmia classification and report generation.
- We release PTB-XL+X, an expert-validated benchmark dataset for evaluating anomaly localization in ECG signals.

2 Related Work

2.1 Foundation models for ECG

Foundation models have become a dominant approach for ECG analysis, typically pretrained on large-scale datasets and adapted to downstream tasks through transfer learning (McKeen et al., 2025; Wang et al., 2025; Song et al., 2025; Mehari and Strodthoff, 2022). However, existing approaches focus primarily on performance without addressing explainability.

ECG-Chat (Zhao et al., 2025) and TolerantECG (Nguyen et al., 2025) adopt multimodal alignment frameworks to align ECG signals with text reports, addressing training noise from similar reports through signal interval features and feature retrieval methods respectively. In a different study, ST-MEM (Na et al., 2024) adopts self-supervised learning through masked patch reconstruction, learning joint spatio-temporal representations that adapt to varying lead configurations. ECG-FM (McKeen et al., 2025) combines contrastive learning with signal masking using domain-specific augmentations.

While these models achieve strong classification performance, none provide interpretable localization of abnormal regions. This gap limits their clinical utility, as physicians cannot verify which ECG features contributed to a diagnosis.

2.2 Explainable AI

Explainable AI aims to enhance transparency in deep learning systems, a critical need in the medical domain where physicians must understand a model's reasoning before trusting its predictions.

Grad-CAM (Selvaraju et al., 2019) produces heatmaps highlighting influential input regions for CNNs. For Transformer architectures, attention calibration (Lu et al., 2021; Zhou et al., 2024) introduces auxiliary losses guided by predefined attention labels to shape attention distributions towards interpretable patterns.

However, attention calibration requires ground-truth attention annotations, which do not exist for published ECG datasets. Our work addresses this annotation bottleneck by automatically generating pseudo-labels from clinical heuristics, enabling attention calibration without manual labeling.

3 Model Architecture

As illustrated in Figure 2, X-ECG consists of three main components: the ECG Encoder, ECG-Text Alignment, and CGAL. Given a 12-lead ECG signal $\mathbf{X} \in \mathbb{R}^{12\times5000}$ and its corresponding text report $\mathbf{R}$, X-ECG learns multimodal representations while simultaneously learning to focus on clinically relevant regions.

The ECG Encoder (center) processes the raw 12-lead ECG through patching, Spatial-Temporal (ST) embedding (Section 3.1), and a Vision Transformer to produce ECG representations. The CLS token aggregates global information, and its attention scores over the ECG patches form the basis for explainability. ECG-Text Alignment (left) aligns ECG representations with text reports preprocessed by the Criteria Feature Retrieval (CFR) module, using a frozen Text Encoder and a Multimodal Text Decoder optimized through contrastive ($\mathcal{L}_{contrast}$) and captioning ($\mathcal{L}_{caption}$) losses (Section 3.2). CGAL (right) addresses the annotation bottleneck by automatically generating attention pseudo-labels: UNet3+ segments ECG waveforms, Abnormal Seeking identifies irregular regions using clinical heuristics, and CFR refines these into pseudo-labels. The attention loss ($\mathcal{L}_{attn}$) then guides the CLS attention toward these clinically meaningful regions (Section 3.3).

3.1 Spatial-Temporal Embedding

Given a raw 12-lead ECG signal $\mathbf{X} \in \mathbb{R}^{L\times T}$, where $L = 12$ denotes the number of leads and

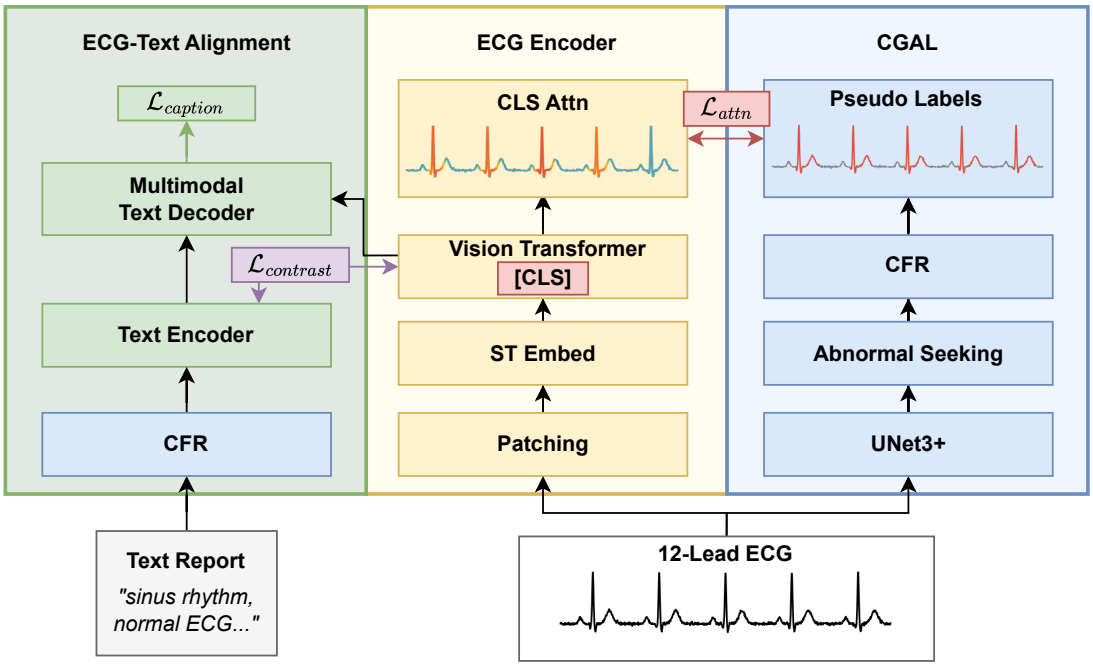


Figure 2: X-ECG overview. The ECG Encoder learns representations through ECG-Text Alignment while CGAL generates pseudo-labels to supervise attention toward clinically relevant regions.

$T = 5000$ is the temporal length (10 seconds at 500Hz), we first apply a linear patch embedding layer. Each lead is divided into $P = T/s = 100$ non-overlapping patches of size $s = 50$, yielding patch embeddings $\mathbf{Z}^{(l)} \in \mathbb{R}^{P\times D}$ for each lead $l \in \{0, \ldots, L-1\}$, where D is the hidden dimension. We then flatten across leads by concatenating all patches sequentially (I, II, III, aVR, aVL, aVF, V1–V6):

$$\mathbf{Z} = [\mathbf{Z}^{(0)}; \mathbf{Z}^{(1)}; \ldots; \mathbf{Z}^{(L-1)}] \in \mathbb{R}^{N\times D}, \quad (1)$$

where $N = L \times P = 1200$ is the number of tokens.

In addition to the Transformer's positional embedding $\mathbf{E} \in \mathbb{R}^{N\times D}$, we incorporate the Spatial-Temporal (ST) embedding proposed by Jin et al. (2025) to enhance the model's ability to capture lead-specific and temporal information, as illustrated in Figure 3. For spatial embedding, we define $\mathbf{E}^s \in \mathbb{R}^{L\times D}$, assigning a unique embedding to each lead (e.g., $E^s_{\mathrm{I}}, E^s_{\mathrm{II}}, \ldots, E^s_{\mathrm{V6}}$). For temporal embedding, we define $\mathbf{E}^t \in \mathbb{R}^{P\times D}$, encoding the position within each lead (i.e., $E^t_0, E^t_1, \ldots, E^t_{P-1}$). The final input to the Transformer is:

$$\mathbf{H}^{(0)}_n = \mathbf{Z}_n + \mathbf{E}_n + \mathbf{E}^s_{l(n)} + \mathbf{E}^t_{t(n)} \quad (2)$$

where $l(n) = \lfloor n/P \rfloor$ and $t(n) = n \bmod P$ for $n = 0, \ldots, N-1$. Here $\mathbf{H}^{(0)}_n \in \mathbb{R}^D$ is the n-th token's input representation (0-indexed), combining its patch embedding with positional embedding

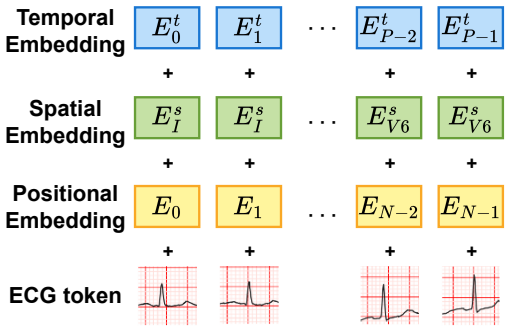


Figure 3: Each ECG token is enriched with Spatial and Temporal embeddings in addition to Positional embeddings, providing explicit lead identity and time frame information.

$(E_0, E_1, \ldots, E_{N-1})$, spatial embedding, and temporal embedding.

3.2 ECG Signal and Text Report Alignment

Once the ECG tokens are enriched with spatial-temporal information, they are passed through the ECG Encoder, a standard Vision Transformer with K layers. A learnable CLS token $\mathbf{h}^{(0)}_{cls} \in \mathbb{R}^D$ is prepended to the sequence:

$$\tilde{\mathbf{H}}^{(0)} = [\mathbf{h}^{(0)}_{cls}; \mathbf{H}^{(0)}] \in \mathbb{R}^{(N+1)\times D}. \quad (3)$$

The encoder applies K transformer layers with multi-head self-attention (MHSA) and feed-

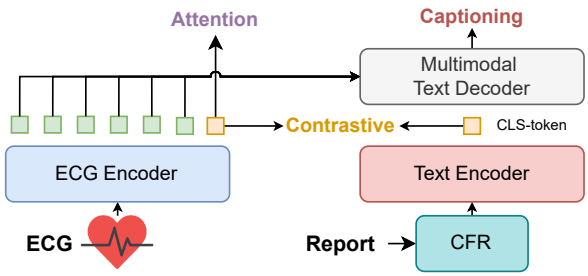


Figure 4: ECG-Text Alignment module. The ECG encoder processes waveforms while the frozen text encoder handles reports preprocessed by CFR. Contrastive learning aligns CLS tokens across modalities, and the multimodal text decoder reconstructs reports conditioned on ECG embeddings.

forward networks (FFN):

$$\tilde{\mathbf{H}}^{(k)} = \text{FFN}(\text{MHSA}(\tilde{\mathbf{H}}^{(k-1)})), \quad k = 1, \ldots, K. \tag{4}$$

The final CLS token $\mathbf{h}_{cls}^{(K)} \in \mathbb{R}^D$ serves as the global ECG representation $\mathbf{E}$ for downstream tasks.

Following previous ECG pretraining approaches (Yu et al., 2024; Zhao et al., 2025), as shown in Figure 4, we employ two pretraining objectives: contrastive loss for representation learning and captioning loss for semantic alignment. Before encoding, text reports are processed by the Criteria Feature Retrieval (CFR) module (Nguyen et al., 2025), which retrieves relevant diagnostic criteria to reduce noise from ambiguous or similar reports. Following CoCa (Yu et al., 2022), the Multimodal Text Decoder employs causal self-attention and cross-attention over ECG encoder outputs for autoregressive report generation. The ECG Encoder and Multimodal Text Decoder are jointly optimized while the Text Encoder remains frozen.

Optimizing the contrastive loss enables alignment between ECG features $\mathbf{E} \in \mathbb{R}^D$ and corresponding text embeddings $\mathbf{T} \in \mathbb{R}^D$. For a batch of B samples, let $\{\mathbf{E}_i, \mathbf{T}_i\}_{i=1}^B$ denote the paired ECG and text representations, and τ the temperature parameter:

$$\mathcal{L}_{contrast} = -\frac{1}{2B}\left(\sum_{i=1}^{B} \log \frac{e^{\mathbf{E}_i^\top \mathbf{T}_i/\tau}}{\sum_{j=1}^{B} e^{\mathbf{E}_i^\top \mathbf{T}_j/\tau}} + \sum_{i=1}^{B} \log \frac{e^{\mathbf{T}_i^\top \mathbf{E}_i/\tau}}{\sum_{j=1}^{B} e^{\mathbf{T}_i^\top \mathbf{E}_j/\tau}}\right) \tag{5}$$

Optimizing the captioning loss encourages the ECG encoder to produce representations that enable accurate report generation. Given the ECG token sequence $\mathbf{H}^{(K)} \in \mathbb{R}^{(N)\times D}$ from the encoder (excluding CLS token) and target text tokens $\mathbf{t} = (t_1, \ldots, t_M)$ of length M, the Multimodal Text Decoder with parameters θ autoregressively predicts each token:

$$\mathcal{L}_{caption} = -\sum_{m=1}^{M} \log P_\theta(t_m | t_{1:m-1}, \mathbf{H}^{(K)}) \tag{6}$$

Algorithm 1 CGAL Pseudo-Label Generation

Require: ECG signal $\mathbf{X}$, diagnosis d
Ensure: Binary attention mask $\mathbf{m} \in \{0, 1\}^N$

// Stage 1: Abnormal Seeking
$\mathbf{S} \leftarrow$ UNet3+($\mathbf{X}$) {segment into P, QRS, T}
(PR, QTc, ST) $\leftarrow$ ComputeMetrics($\mathbf{S}$)
$\mathcal{R} \leftarrow$ FlagAbnormal(PR, QTc, ST)
// Stage 2: Criteria Seeking
$criteria \leftarrow$ CFR(d)
$\mathcal{R} \leftarrow \mathcal{R} \cap criteria$ {keep diagnosis-relevant}
$\mathcal{R} \leftarrow \mathcal{R} \cup$ SinusRegions($\mathbf{S}$) {add P, QRS in leads II, aVR}
$\mathbf{m} \leftarrow$ CreateMask($\mathcal{R}, N$)
return $\mathbf{m}$

3.3 Clinically-Guided Attention Localization (CGAL)

CGAL is the key component that solves the annotation bottleneck. It generates attention pseudo-labels on-the-fly during training by encoding established clinical knowledge into a rule-based algorithm. The procedure is summarized in Algorithm 1.

The first stage (*Abnormal Seeking*) identifies candidate abnormal regions using clinical definitions from the *Life in the Fastlane* ECG library (lit, 2008). A pretrained UNet3+ model (Huang et al., 2020), trained on SemiSegECG (Park et al., 2025) and RDB (Liu et al., 2025) datasets and frozen during our training, produces a segmentation map $\mathbf{S}^{(l)} \in \{0, 1, 2, 3\}^T$ for each lead l, where labels correspond to background (0), P wave (1), QRS complex (2), and T wave (3). From these segment boundaries, ComputeMetrics() computes clinical metrics: PR interval (P onset to QRS onset), QT interval with Bazett's correction, and ST segment deviation. FlagAbnormal() then flags regions violating established thresholds (Appendix A). For QT analysis, we apply:

$$QTc = \frac{QT}{\sqrt{RR}}, \tag{7}$$

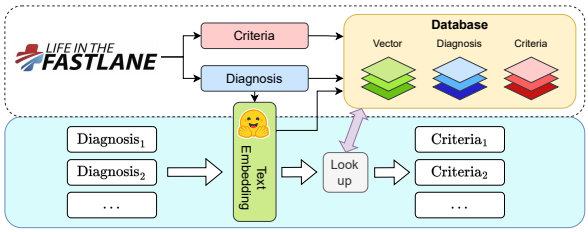


Figure 5: The Criteria Feature Retrieval (CFR) module retrieves diagnostic criteria by comparing the input diagnosis embedding with a pre-constructed vector database.

where QT is the measured interval and RR is the preceding R-R interval. This correction accounts for the physiological shortening of QT at higher heart rates.

The second stage (*Criteria Seeking*) filters these candidates using CFR() (Nguyen et al., 2025), illustrated in Figure 5. CFR() maintains a vector database $\mathcal{D} = \{(\mathbf{d}_j, c_j)\}_{j=1}^{|\mathcal{D}|}$ of diagnosis-criteria pairs sourced from the LITFL ECG library, where $\mathbf{d}_j \in \mathbb{R}^D$ is the embedding of diagnosis j and c_j is its defining criteria (e.g., "prolonged PR interval $>$ 200ms" for first-degree AV block). Given an input diagnosis with embedding $\mathbf{q} \in \mathbb{R}^D$, CFR() retrieves the most similar criteria via cosine similarity:

$$j^* = \text{argmax}_j \frac{\mathbf{q}^\top \mathbf{d}_j}{\|\mathbf{q}\|\|\mathbf{d}_j\|} \tag{8}$$

We retain only flagged regions matching these retrieved criteria: for instance, keeping prolonged PR intervals for first-degree AV block while discarding unrelated ST findings. When no criteria match (e.g., for rare diagnoses not in our database), attention supervision is skipped for that sample and training proceeds with only $\mathcal{L}_{contrast}$ and $\mathcal{L}_{caption}$. Since this pipeline only captures abnormalities, SinusRegions() explicitly adds P wave and QRS markers in leads II and aVR as sinus rhythm reference points. Lead II is chosen because the heart's electrical axis aligns with this lead, producing the tallest and clearest P waves, making it the standard rhythm strip in clinical practice. Lead aVR provides a unique rightward-superior viewpoint where P waves are normally inverted; upright P waves in aVR indicate lead misplacement or ectopic atrial rhythms, making it essential for confirming normal sinus origin. Finally, CreateMask() converts these regions into a binary token mask $\mathbf{m} \in \{0,1\}^N$ by setting $m_n = 1$ for tokens overlapping with any flagged region in $\mathcal{R}$, and $m_n = 0$ otherwise.

Attention Supervision. From the preceding stages, we obtain a binary mask $\mathbf{m}_i \in \{0,1\}^N$ for sample i, where $m_{i,n} = 1$ indicates an abnormal region at token n. We normalize this mask into a target probability distribution $\mathbf{p}_i \in \mathbb{R}^N$:

$$p_{i,n} = \frac{m_{i,n}}{\sum_{j=0}^{N-1} m_{i,j}}, \quad n = 0, \ldots, N-1 \tag{9}$$

To extract the model's attention pattern, we compute the attention weights from the CLS token in the final transformer layer. For attention head h, let $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{(N+1)\times d_h}$ denote the query and key matrices, where $d_h = D/H$ is the head dimension and H is the number of heads. The CLS token's attention over the N ECG tokens is:

$$\mathbf{a}^{(h)} = \text{softmax}\left(\frac{\mathbf{K}_{1:N}\mathbf{q}_{cls}}{\sqrt{d_h}}\right) \in \mathbb{R}^N, \tag{10}$$

where $\mathbf{q}_{cls} = \mathbf{Q}_0 \in \mathbb{R}^{d_h}$ is the CLS query vector (first row of $\mathbf{Q}$) and $\mathbf{K}_{1:N} \in \mathbb{R}^{N\times d_h}$ contains the keys for ECG tokens. We average across all heads to obtain the final attention distribution $\mathbf{A}_i^{CLS} = \frac{1}{H}\sum_{h=1}^{H} \mathbf{a}^{(h)} \in \mathbb{R}^N$. The attention loss aligns this distribution with the pseudo-label via KL divergence:

$$\begin{aligned}\mathcal{L}_{attn} &= \frac{1}{B}\sum_{i=1}^{B} D_{KL}(\mathbf{p}_i \| \mathbf{A}_i^{CLS}) \\ &= \frac{1}{B}\sum_{i=1}^{B}\sum_{n=0}^{N-1} p_{i,n} \log \frac{p_{i,n}}{A_{i,n}^{CLS}}\end{aligned} \tag{11}$$

To preserve general dependencies, only the final layer's attention is supervised; early layers remain unconstrained.

3.4 Training Objective

The overall training loss combines all objectives:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{contrast} + \beta\mathcal{L}_{caption} + \gamma\mathcal{L}_{attn} \tag{12}$$

where α, β and γ control the relative contributions of contrastive learning, captioning, and attention supervision, respectively.

Through this composite objective, contrastive loss aligns ECG features with textual representations, captioning loss enables accurate generation of clinical descriptions, and attention-guiding loss regularizes the model to focus on medically relevant regions.

4 Experiment

4.1 Dataset

We pretrain on MIMIC-IV-ECG (Gow et al., 2023), comprising over 800,000 12-lead ECG recordings (10-second duration) from approximately 160,000 patients. During preprocessing, we replace NaN/Inf values with zero. CGAL generates attention pseudo-labels on-the-fly without external annotations.

For evaluation, we employ the PTB-XL dataset (Strodthoff et al., 2020), a clinical 12-lead ECG dataset containing 21,799 recordings from 18,869 patients. To create `PTB-XL+X`, an expert-validated benchmark for anomaly localization, we applied the CGAL pipeline to generate candidate heatmaps for all 2,198 test recordings, which were then reviewed and corrected by board-certified cardiologists to produce ground truth annotations (see Appendix B for annotation protocol and samples). Importantly, these expert annotations serve as evaluation ground truth, distinct from the automatically generated pseudo-labels used during training. For classification, we evaluate on the PTB-XL All task, which includes all diagnostic categories. For report generation, we adopt the LLM pretraining and fine-tuning protocols from ECG-Chat (Zhao et al., 2025).

4.2 Configuration

We use a 1D 12-layer Vision Transformer with patch size 50 as the ECG Encoder. The text encoder (MedCPT (Jin et al., 2023)) remains frozen. We flatten the 12-lead ECG signal to ensure the CLS token attends to all tokens. Training uses learning rate $1e-4$, 20 epochs, batch size 64 on $4\times$ A100 GPUs. We set $\alpha = 1$, $\beta = 2$ (following (Zhao et al., 2025)), and$\gamma = 0.5$. To reduce computational overhead, CGAL pseudo-labels are cached during training.

For fair comparison, we retrained TolerantECG (Nguyen et al., 2025) using the same encoder and configuration. Since TolerantECG performs double steps per iteration, we trained it for half the epochs.

4.3 Metrics

Anomaly localization. To evaluate whether the attention scores of `X-ECG` accurately highlight relevant regions within each data sample, we frame this as a ranking task where the model should assign higher attention to abnormal tokens. We employ Precision at K (P@K) and Normalized Discounted Cumulative Gain at K (NDCG@K) where K = 50% of the total token count (K=600 for N=1200), as the large sequence length makes fixed small K values less discriminative. We also report point-level Area Under the Receiver Operating Characteristic Curve (AUROC), following the protocol in (Jiang et al., 2023b).

Arrhythmia Classification. We adopt macro-averaged AUROC, which computes the mean performance across all classes regardless of class imbalance.

For clarity, we denote localization AUROC as $\text{AUROC}_{\text{loc}}$ and classification AUROC as $\text{AUROC}_{\text{cls}}$ throughout the paper.

Report generation. We evaluate the output of the LLM conditioned on ECG representations using ROUGE-L to assess lexical similarity between the generated and reference reports. Additionally, we employ BERT-score to assess semantic similarity, capturing how closely the generated report aligns with the meaning of the reference report.

5 Results and Analysis

5.1 Anomaly Localization

To evaluate the model's ability to identify components contributing to the final decision output, we perform inference on the PTB-XL test set and benchmark it as a binary point-wise classification task, as shown in Table 1. Ground truth labels come from the expert-validated annotations in `PTB-XL+X`, ensuring evaluation independence from the training pseudo-labels.

`X-ECG` achieves 84% $\text{AUROC}_{\text{loc}}$ on `PTB-XL+X`, even without exposure to PTB-XL samples during pretraining, demonstrating that our self-supervised attention mechanism learns to localize abnormal regions through clinical heuristics alone. In contrast, the attention scores produced by ECG-Chat exhibit limited reliability with $\text{AUROC}_{\text{loc}}$ close to 50%, indicating nearly random focus. Notably, ECGAD, a model specifically designed for anomaly localization, achieves only 65% $\text{AUROC}_{\text{loc}}$. The other metrics show similar trends: `X-ECG` achieves 35.36% P@K versus 23.70% for ECGAD and 20–21% for other foundation models, and 83.04% NDCG@K versus 60.60% for ECGAD. On the ECGAD dataset, the margins are narrower (72.75% vs. 71.49% $\text{AUROC}_{\text{loc}}$) but `X-ECG` still leads across all metrics.

As shown in Table 1 and Table 2, `X-ECG` is capable of performing both anomaly localization and

Table 1: Anomaly localization results on PTB-XL+X and the abnormal dataset provided by ECGAD (Jiang et al., 2023b). **Bold** indicates the best result, and underline indicates the second-best result.

Methods	PTB-XL+X			ECGAD		
	P@50	NDCG@50	$\text{AUROC}_{\text{loc}}$	P@50	NDCG@50	$\text{AUROC}_{\text{loc}}$
ECGAD (Jiang et al., 2023b)	23.70	60.60	65.56	22.26	67.00	71.49
ST-MEM (Na et al., 2024)	20.72	43.61	50.00	16.29	40.63	50.36
ECG-FM (McKeen et al., 2025)	20.34	42.68	49.41	15.99	40.24	49.26
ECG-Chat (Zhao et al., 2025)	20.87	44.55	50.64	16.33	41.81	49.78
TolerantECG (Nguyen et al., 2025)	21.10	47.31	52.99	16.73	46.06	53.13
X-ECG (Ours)	**35.36**	**83.04**	**84.19**	**24.08**	**67.47**	**72.75**

Table 2: Linear probing classification result evaluated on PTB-XL dataset. **Bold** indicates the best result, and underline indicates the second-best result

Methods	$\text{AUROC}_{\text{cls}}$
ST-MEM (Na et al., 2024)	87.16
ECG-FM (McKeen et al., 2025)	85.20
TolerantECG (Nguyen et al., 2025)	**92.11**
X-ECG (Ours)	91.72

Table 3: Report generation evaluation using the English-translated PTB-XL report as label

Methods	ROUGE-L	BERT-score
ECG-Chat (Zhao et al., 2025)	33.83	89.00
X-ECG (Ours)	**33.97**	**89.12**

Table 4: Effect of attention supervision on localization and classification.

Configuration	$\text{AUROC}_{\text{loc}}$	$\text{AUROC}_{\text{cls}}$
$\gamma = 0$ (no attention-guiding)	46.40	91.38
$\gamma = 1.0$	83.86	91.22
X-ECG ($\gamma = 0.5$)	**84.19**	**91.72**

arrhythmia classification, whereas ECGAD is limited to identifying abnormal regions without the ability to classify the underlying diagnosis.

Figure 6 presents qualitative comparisons of anomaly localization. In example (a), ECGAD highlights redundant QRS complexes rather than the diagnostically relevant inverted T wave, whereas X-ECG correctly identifies the abnormal region matching expert annotation. In example (b), ECGAD completely misses the absent P waves characteristic of atrial fibrillation, while X-ECG accurately localizes them. These cases illustrate how X-ECG's clinically-guided attention captures diagnosis-specific abnormalities that reconstruction-based methods overlook.

5.2 Arrhythmia classification

Upon completing pretraining, we attach a linear classification head using the CLS token embedding as input. To evaluate representation quality, we freeze the ECG Encoder and train only the linear head. For ST-MEM and ECG-FM, we use official pretrained weights with the same procedure. Table 2 presents classification results on PTB-XL.

X-ECG achieves competitive classification performance (91.72% $\text{AUROC}_{\text{cls}}$), close to TolerantECG, while uniquely providing anomaly localization capabilities unavailable in other approaches.

5.3 Report Generation

We extend LLaVA (Liu et al., 2023) to the ECG domain by connecting the pretrained ECG Encoder with Vicuna-7B (Zheng et al., 2023). A two-layer MLP projection transforms ECG embeddings into the LLM's text embedding space. During initial training, both encoder and LLM are frozen while the projection layer is optimized. Subsequently, we fine-tune the LLM using LoRA (Hu et al., 2021). Table 3 compares report generation quality against ECG-Chat (Zhao et al., 2025). X-ECG outperforms on both ROUGE-L and BERT-score, indicating improved lexical precision and semantic fidelity.

5.4 Ablation Study

Attention Supervision. Table 4 shows the impact of our attention-guiding mechanism. Without it, localization drops to 46.40% (near random), while classification remains stable at 91.38%. This confirms that attention-guiding is essential for explainability but does not harm discriminative learning. Increasing the attention weight ($\gamma = 1.0$) causes slight degradation in both tasks, suggesting that over-constraining attention can hurt generalization.

Patch Size. Table 5 compares patch sizes. Larger patches (100 vs. 50) reduce both localiza-

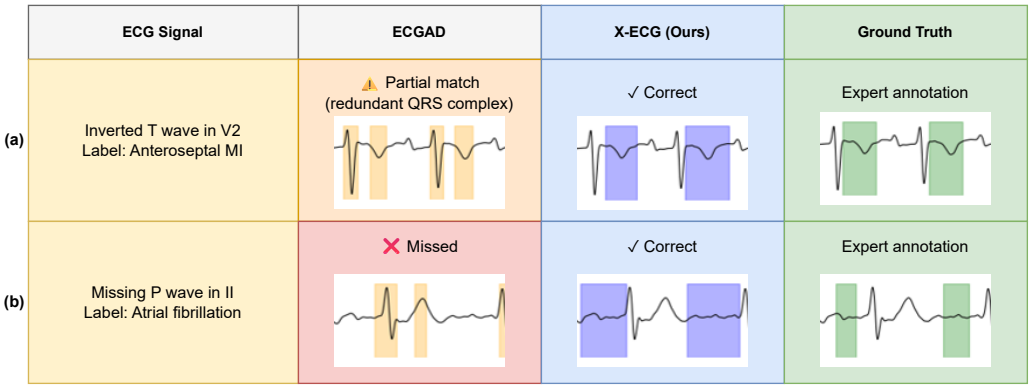


Figure 6: Qualitative comparison with other methods on anomaly localization. Each row shows an ECG case with the original signal (left), localization results from ECGAD and X-ECG (middle), and expert annotation (right). Highlighted regions indicate detected abnormalities: (a) inverted T wave in V2 for anteroseptal MI, (b) missing P wave in lead II for atrial fibrillation.

Table 5: Effect of patch size.

Patch Size	$\text{AUROC}_{\text{loc}}$	$\text{AUROC}_{\text{cls}}$
100	78.57	90.06
50 (X-ECG)	**84.19**	**91.72**

Table 6: Effect of Spatial-Temporal embedding.

Configuration	$\text{AUROC}_{\text{loc}}$	$\text{AUROC}_{\text{cls}}$
w/o ST embedding	81.24	**91.82**
w/ ST embedding (X-ECG)	**84.19**	91.72

tion and classification, likely due to coarser temporal resolution that prevents precise localization of abnormal regions.

Spatial-Temporal Embedding. Table 6 examines the ST embedding contribution. Removing it yields a marginal classification gain (+0.1%) but substantially degrades localization (−2.95%), indicating that explicit lead and time information helps the model attend to the correct regions. We retain ST embedding as the localization improvement outweighs the minor classification trade-off.

6 Conclusion

We introduced X-ECG, an explainable ECG foundation model that addresses a fundamental limitation of existing approaches: the inability to show clinicians where abnormalities occur. Our key insight is that established clinical knowledge can automatically generate attention supervision signals, transforming the explainability problem from an annotation bottleneck into an engineering task.

Through CGAL, X-ECG learns to focus on clinically relevant regions without manual annotation, achieving high result on anomaly localization while achieving state-of-the-art classification performance. When X-ECG predicts a diagnosis, it highlights the specific ECG waves that contributed to that decision, enabling clinicians to verify model reasoning in seconds. We release PTB-XL+X, an expert-validated benchmark for anomaly localization, to facilitate further research in explainable ECG analysis.

Limitations

The CGAL pseudo-label generation relies on established clinical knowledge to identify abnormal conditions during training. However, this approach may overlook certain cases due to dependencies on patient-specific factors such as gender and age. While the evaluation benchmark PTB-XL+X uses expert-validated annotations, the training signal remains limited by the coverage of our rule-based heuristics. Future work could incorporate more comprehensive clinical guidelines or learn condition-specific attention patterns to improve pseudo-label quality.

Ethics Statement

This work uses publicly available datasets (MIMIC-IV-ECG and PTB-XL) that have undergone institutional de-identification procedures. No personally identifiable information was accessed during this research. X-ECG is designed as a clinical decision support tool, not a replacement for physician judgment. All predictions require verification by qualified healthcare professionals before clinical action.

References

2008. Life in the fast lane.

Nhat-Tan Bui, Dinh-Hieu Hoang, Thinh Phan, Minh-Triet Tran, Brijesh Patel, Donald Adjeroh, and Ngan Le. 2024. Tsrnet: Simple framework for real-time ecg anomaly detection with multimodal time and spectrogram restoration network. *Preprint*, arXiv:2312.10187.

Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. 2023. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset (version 1.0).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. 2020. Unet 3+: A full-scale connected unet for medical image segmentation. *Preprint*, arXiv:2004.08790.

Aofan Jiang, Chaoqin Huang, Qing Cao, Shuang Wu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng Wang. 2023a. Multi-scale cross-restoration framework for electrocardiogram anomaly detection. *Preprint*, arXiv:2308.01639.

Aofan Jiang, Chaoqin Huang, Qing Cao, Shuang Wu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng Wang. 2023b. Multi-scale cross-restoration framework for electrocardiogram anomaly detection. *Preprint*, arXiv:2308.01639.

Aofan Jiang, Chaoqin Huang, Qing Cao, Yuchen Xu, Zi Zeng, Kang Chen, Ya Zhang, and Yanfeng Wang. 2024. Anomaly detection in electrocardiograms: Advancing clinical diagnosis through self-supervised learning. *Preprint*, arXiv:2404.04935.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. 2025. Reading your heart: Learning ecg words and sentences via pre-training ecg language model. *Preprint*, arXiv:2502.10707.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Yuhang Liu, Peng Zhang, and Fan Lin. 2025. Resting ECG Segmentation Dataset.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2021. Attention calibration for transformer in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1288–1298, Online. Association for Computational Linguistics.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. 2025. Ecg-fm: An open electrocardiogram foundation model. *Preprint*, arXiv:2408.05178.

Temesgen Mehari and Nils Strodthoff. 2022. Self-supervised representation learning from 12-lead ecg data. *Computers in Biology and Medicine*, 141:105114.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. 2024. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *Preprint*, arXiv:2402.09450.

Huynh Dang Nguyen, Trong-Thang Pham, Ngan Le, and Van Nguyen. 2025. Tolerantecg: A foundation model for imperfect electrocardiogram. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 8097–8105, New York, NY, USA. Association for Computing Machinery.

Minje Park, Jeonghwa Lim, Taehyung Yu, and Sunghoon Joo. 2025. Semisegecg: A multi-dataset benchmark for semi-supervised semantic segmentation in ecg delineation. *Preprint*, arXiv:2507.18323.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Junho Song, Jong-Hwan Jang, DongGyun Hong, Joon myoung Kwon, and Yong-Yeon Jo. 2025. Crema: A contrastive regularized masked autoencoder for robust ecg diagnostics across clinical domains. *Preprint*, arXiv:2407.07110.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. 2020. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *Preprint*, arXiv:2004.13701.

Yue Wang, Xu Cao, Yaojun Hu, Haochao Ying, Hongxia Xu, Ruijia Wu, James Matthew Rehg, Jimeng Sun, Jian Wu, and Jintai Chen. 2025. Anyecg: Foundational models for multitask cardiac analysis in real-world settings. *Preprint*, arXiv:2411.17711.

Han Yu, Peikun Guo, and Akane Sano. 2024. Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text. *Preprint*, arXiv:2405.19366.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Preprint*, arXiv:2205.01917.

Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. 2025. Ecg-chat: A large ecg-language model for cardiac disease diagnosis. *Preprint*, arXiv:2408.08849.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Peilin Zhou, Qichen Ye, Yueqi Xie, Jingqi Gao, Shoujin Wang, Jae Boum Kim, Chenyu You, and Sunghun Kim. 2024. Attention calibration for transformer-based sequential recommendation. *Preprint*, arXiv:2308.09419.

A Abnormal criteria in ECG signal

In the first stage of the disease heatmap curation pipeline, we identify and highlight abnormal regions across individual waves and leads. To achieve this, we collect diagnostic conditions from LITFL (lit, 2008) and subsequently verify them with clinical experts to ensure medical validity. These curated conditions serve as the foundation for constructing reliable heatmaps that emphasize clinically relevant abnormalities. A summary of the identified conditions is presented in Table 7.

Components	Abnormal conditions
P wave	• Missing • Amplitude $>$ 0.25mV in limb leads • Amplitude $>$ 0.15mV in precordial leads • Duration $>$ 120ms • Inverted in lead I, II, III • Upright in lead aVR
QRS complex	• Duration $>$ 120ms • Has odd shape (RSR', QR, rS)
T wave	• Amplitude $>$ 0.5mV in limb leads • Amplitude $>$ 1mV in precordial leads • Inverted in leads where normally upright (I, II, V2-V6)
PR interval	• Duration $>$ 200ms • Duration $<$ 120ms
QT interval	• Corrected duration $>$ 440ms • Corrected duration $<$ 350ms
ST segment	• Elevate: $>$ 0.1mV isoelectric line ($>$ 0.05mV in lead V2 and V3) • Depression: $<$ 0.5mV isoelectric line

Table 7: Abnormal conditions for each component in ECG signal

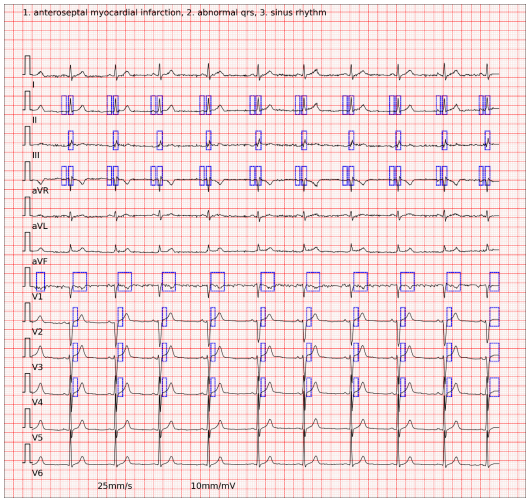
B Annotation Protocol for PTB-XL+X

To construct PTB-XL+X, we employed a two-stage annotation protocol. First, the CGAL pipeline generated candidate heatmaps highlighting potential abnormal regions. Second, board-certified cardiologists reviewed each candidate heatmap, correcting false positives, adding missed abnormalities, and verifying alignment with clinical diagnostic criteria. This expert review ensures that evaluation ground truth reflects clinical judgment rather than algorithmic heuristics.

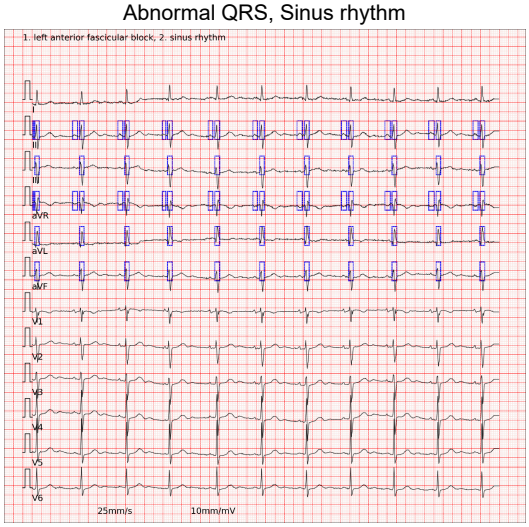
Figure 7 showcases representative examples of the final expert-validated annotations alongside their corresponding original text reports from PTB-XL+X.

C Details of Large Language Models Usage

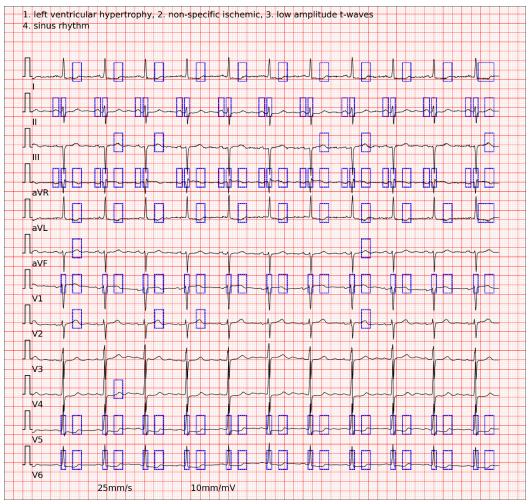
The use of Large Language Models (LLMs) in this work is limited solely to grammar correction and stylistic refinement. All core aspects—including the formulation of main contributions, experimental design, and data analysis—were conducted independently without LLM involvement.



Anteroseptal myocardial infraction,
Abnormal QRS, Sinus rhythm



Left anterior fascicular block, Sinus rhythm



Left ventricular hypertrophy, Non-specific ischemic,
Low amplitude T-wave, Sinus rhythm

Figure 7: Example expert-validated annotations from PTB-XL+X. Blue boxes indicate clinically relevant regions confirmed by cardiologists.